Transformações nos Dados

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte

Introdução

Introdução

- Nem sempre os dados que desejamos analisar vem preparados para isto
- Boa parte dos algoritmos que iremos ver no curso terão desempenho melhor se aplicados em dados simétricos e com variância unitária
- Portanto, muitas vezes é necessário pré-processar os dados antes de começar nossa análise

Transformações nos Dados

Centering e Scaling

- São as duas transformações mais comuns
- Centering significa subtrair a média dos dados
- Scaling envolve dividir os dados pelo seu desvio padrão
- Ou seja, esta transformação nada mais é do que transformar todo x_i em um x_i* tal que

$$x_i* = \frac{x_i - \bar{x}}{S_x}$$

Centering e Scaling

- Devido a esta transformação, os dados ficam com média 0 e desvio padrão 1
- Esta transformação é muito utilizada quando as variáveis estão em escala diferentes
- Entretanto, perdemos interpretabilidade nos dados individuais

Centering e Scaling

- Felizmente o `R` é capaz de resolver facilmente este problema
- A função `scale` calcula automaticamente a média e o desvio padrão das colunas dos conjuntos de dados e aplica a transformação desejada

Centering e Scaling

```
> library(dplyr)
> iris %>%
+   select(-Species) %>%
+   summarise_all(c("mean", "sd"))
```

```
##   Sepal.Length_mean Sepal.Width_mean Petal.Length_mean
## 1          5.843333         3.057333             3.758
##   Petal.Width_mean Sepal.Length_sd Sepal.Width_sd
## 1         1.199333       0.8280661      0.4358663
##   Petal.Length_sd Petal.Width_sd
## 1        1.765298      0.7622377
```

Centering e Scaling

```
> iris_cs <- as.data.frame(scale(iris[, -5]))
> iris_cs %>%
+   summarise_all(c("mean", "sd"))

##   Sepal.Length_mean Sepal.Width_mean Petal.Length_mean
## 1     -4.484318e-16     2.034094e-16     -2.895326e-17
##   Petal.Width_mean Sepal.Length_sd Sepal.Width_sd
## 1    -3.663049e-17               1              1
##   Petal.Length_sd Petal.Width_sd
## 1               1              1
```

Transformações para Eliminar Assimetria

- A assimetria amostral de um vetor de dados $\boldsymbol{x} = x_1, \cdots, x_n$ é dada por

$$b_1 = \frac{\sum (x_i - \bar{x})^3}{(n-1)\nu^{3/2}},$$

em que

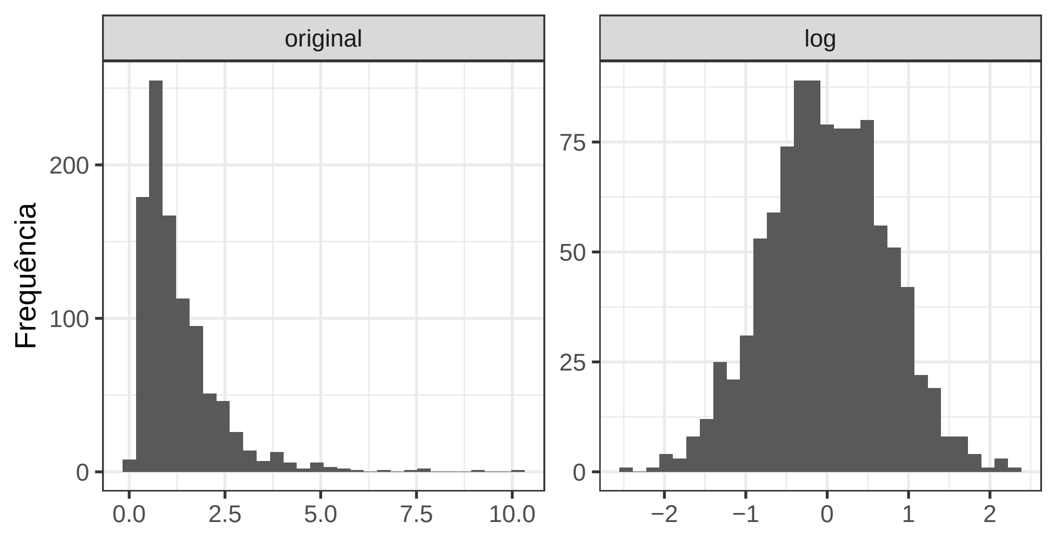
$$\nu = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

- Se $b_1 = 0$, então a distribuição é simétrica
- Se $b_1 > 0$, então a distribuição é assimétrica à direita
- Se $b_1 < 0$, então a distribuição é assimétrica à esquerda

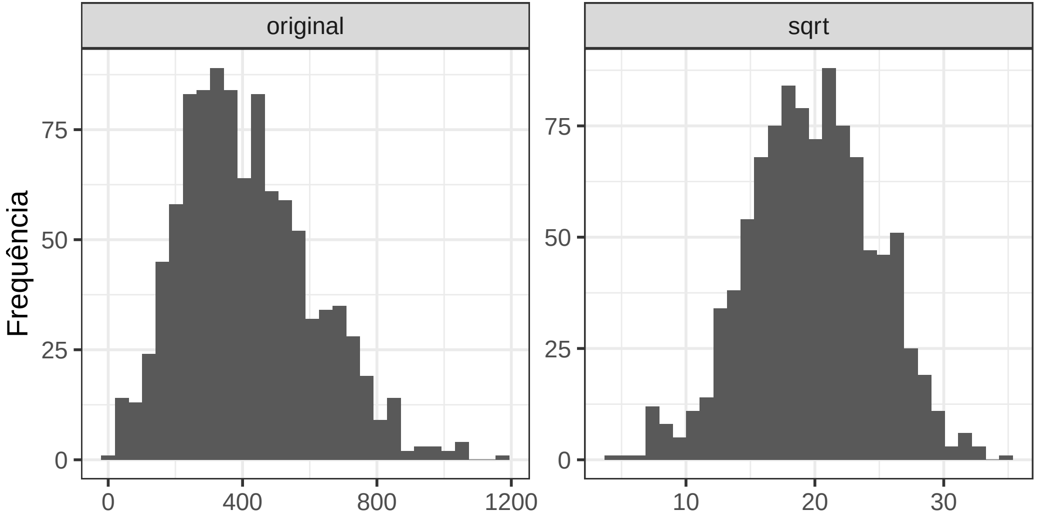
Transformações para Eliminar Assimetria

- Algumas transformações que podem ajudar a eliminar a assimetria são $\log$, raiz quadrada e a inversa
- Os próximos slides mostram algumas destas transformações na prática

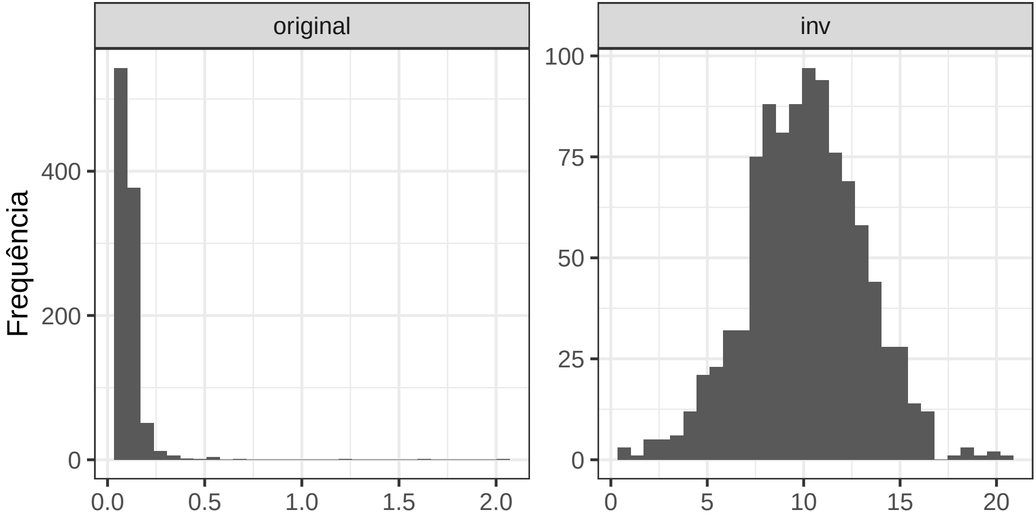
Transformações para Eliminar Assimetria - log



Transformações para Eliminar Assimetria - Raiz Quadrada



Transformações para Eliminar Assimetria - Inversa



Transformação de Box-Cox

- É uma família de transformações definida pela expressão

$$x_i* = \begin{cases} \frac{x_i^{\lambda}-1}{\lambda}, & \text{se } \lambda \neq 0 \\ \log(x_i), & \text{se } \lambda = 0 \end{cases}$$

- É possível utilizar máxima verosimilhança para estimar o melhor valor para λ

Análise de Componentes Principais

- Baseada em autovalores e autovetores
- Permite transformar as variáveis preditoras correlacionadas e obter variáveis não-correlacionadas
- Veremos em detalhes na próxima aula

Variáveis Dummy

Variáveis Dummy

- Imagine que desejamos fazer uma regressão linear
- Para isto, as variáveis resposta e preditora devem ambas serem quantitativas
- Mas e se quisermos fazer uma regressão com variáveis preditoras que sejam qualitativas?

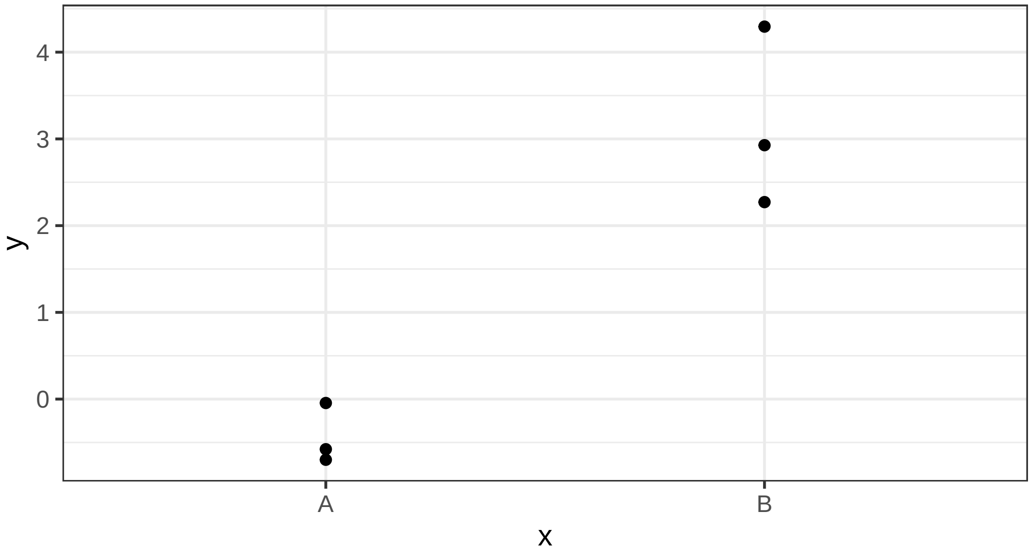
Variáveis Dummy

- Para isto, usamos variáveis dummy (ou variáveis indicadoras)
- Estas variáveis tomam o valor 1 na presença do nível i da variável categórica e 0 na sua ausência
- Assim, uma variável qualitativa de k níveis torna-se k variáveis diferentes
- A partir desta transformação, a regressão é realizada da mesma maneira

Variáveis Dummy

```
> set.seed(1221)
> n <- 3
> y <- c(rnorm(n, mean=0), rnorm(n, mean=2))
> x <- c(rep("A", n), rep("B", n))
>
> dados <- data.frame(x=x, y=y)
> ggplot(dados, aes(x=x, y=y)) +
+   geom_point()
```

Variáveis Dummy



Variáveis Dummy - Padrão do R

```
##   (Intercept) xB
## 1           1  0
## 2           1  0
## 3           1  0
## 4           1  1
## 5           1  1
## 6           1  1
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$x
## [1] "contr.treatment"
```

Variáveis Dummy - Padrão do R

```
##
## ^^ITwo Sample t-test
##
## data:  y by x
## t = -5.7322, df = 4, p-value = 0.004587
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.351139 -1.858884
## sample estimates:
## mean in group A mean in group B
##      -0.4405516       3.1644602
```

Variáveis Dummy - Padrão do R

```
##
## Call:
## lm(formula = y ~ x, data = dados)
##
## Residuals:
##       1       2       3       4       5       6
## -0.2587 -0.1370  0.3957 -0.2367  1.1300 -0.8933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.4406     0.4447  -0.991  0.37793
## xB            3.6050     0.6289   5.732  0.00459 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Outliers

Outliers

- São observações aberrantes
- Dependem do contexto e da distribuição dos dados
- É muito difícil definir precisamente quando uma observação é um outlier

Outliers

- O primeiro passo é verificar se o ponto suspeito é cientificamente válido
- Em amostras pequenas, eventos raros podem ser amplificados
- Simplesmente retirar outliers não é, necessariamente, a melhor forma de lidar com este problema

Outliers

- Os pontos destacados no boxplot tradicional são outliers, mas em um contexto muito específico
- Pontos que são outliers em um contexto podem não ser em outro
- A definição de outlier vai depender diretamente do modelo que assumimos que os dados possuem
- Felizmente, há muitos métodos estatísticos robustos a outliers

Dados Faltantes

Dados Faltantes

- Primeiro é necessário entender o comportamento dos dados faltantes
- Os dados podem ser estruturalmente faltantes, como o número de crianças às quais um homem deu a luz
- Mas existem casos mais complicados de lidar

Dados Faltantes

- De modo análogo aos outliers, é importante saber o porquê dos dados estarem faltando
- É possível que a altura de uma pessoa esteja faltando no conjunto de dados, mas tenhamos todas as suas outras informações
- Será que esta é uma variável muito importante?

Dados Faltantes

- Talvez não seja se quisermos modelar o crédito desta pessoa
- Talvez seja se quisermos modelar taxa de obesidade
- Cada caso é um caso

Como Lidar com Dados Faltantes?

- Ignorar
- Se o conjunto de dados for grande o suficiente, pode haver redundância nos dados
- Talvez não seja possível fazer isto caso sejam poucos os dados disponíveis

Como Lidar com Dados Faltantes?

- Preencher os dados faltantes manualmente, seja recalculando ou medindo novamente as variáveis
- Nem sempre é possível recoletar dados
- É um processo lento e caro devido ao trabalho manual envolvido

Como Lidar com Dados Faltantes?

- Recategorizar os dados, criando um novo nível chamado “faltante”
- Simples e eficaz
- Não é útil para variáveis quantitativas; afinal, 0 é um valor possível ou um valor faltante?

Como Lidar com Dados Faltantes?

- Substituir o valor faltante com a média, no caso de variáveis quantitativas, ou a moda, no caso de variáveis qualitativas
- Funciona bem para dados representativos
- Teremos problemas se os dados faltantes forem outliers

Como Lidar com Dados Faltantes?

- Prever novos valores através de modelagem estatística
- Depende de definir um modelo estatístico para fazer a imputação
- De modo análogo ao caso dos outliers, tudo depende de nossas informações *a priori*

Transformações nos Dados

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte